������������

1. ��

- ���
  ��/��������/���������������������
- ��“���”����������������/������������������
  �� `conversion.handle_batch` �������� DOCX/HTML/PDF→PDF/�
  ��������������
- ��������<1s ������������������������������
  ���������

2. ����

- **���probe�**���������������������������������
- **�����profiling�**����������/��/������������
- **�����strategy recommendation�**�������������
  `heading_block + length_split`�`sentence_split + sliding_window`
  �`table_batch` ��
- **������������**����� `conversion.handle_batch`���������
  ��������

3. ����

1. **����**�
   - ����PDF/DOCX����/�/�� 1~2 ��������������
   - ����PPTX����/�/������������
   - HTML/Markdown���� 3 ���������������
   - ��/��������/������/����
2. **����**�
   - ���heading ����������/����������������
     ���
   - �������/P90���/����������OCR ��������

    - ����/���������������

3. **����**�

- ����`heading_block + length_split(200~400�)`�
- ������`sentence_split + sliding_window(overlap=10~20%)`�
- �����`table_batch + paragraph_split`�
- ��/���`code/log_block + no_overlap`�
- ����`slide_block + textbox_merge`�

4. **����**����������������������/��/���������

4. ����

- **�����Chain**��
    i. `probe.extract_signals`������������������ `profile`�
    ii. `probe.recommend_strategy`��� `profile` �� `strategy_id` �� ��target_length/overlap/preserve_tables ���
    iii. �������������������������� `conversion.handle_batch` ����
- **������**���/�����������������������/�����

5. ��������

```json
{
  "task_id": "probe-and-recommend-123",
  "profile": {
    "heading_ratio": 0.42,
    "list_ratio": 0.18,
    "table_ratio": 0.05,
    "code_ratio": 0.0,
    "p90_para_len": 220,
    "samples": ["H1: ��...", "��...", "��..."],
    "limits": {"media_slice_supported": false}
  },
  "recommendation": {
    "strategy_id": "heading_block_length_split",
    "params": {
      "target_length": 220,
      "overlap_ratio": 0.15,
      "preserve_tables": true
    },
    "candidates": [
      {"strategy_id": "heading_block_length_split", "score": 0.71},
      {"strategy_id": "sentence_split_sliding", "score": 0.66}
    ],
    "notes": "�������������������"
  }
}
```

6. ������

- `has_headings || list_density>��` → `heading_block + length_split(200~400�)`
- `code_density>�� || log_pattern` → `code/log_block + no_overlap`
- `table_density>��` → `table_batch + paragraph_split`
- `else` → `sentence_split + sliding_window(overlap=15%)`

7. ����

- �����O(k) �/���
    - PDF/DOCX�`k<=6`���/�/��������������������
    ����������
    - PPTX���/�/�������� `slide_index` � `textbox_index`�
    - HTML/MD������������������������
    - ��/������� + ��/������������/������
- �������+���O(n) ��������
    - ���
    ��`heading_ratio`�`list_ratio`�`table_ratio`�`code_ratio`
    �`slide_textbox_cnt`�
    - �����������/��/`p90`�`digit_symbol_ratio`����
    �shingle ������OCR �����/�������
    - ����������������������/�� Jaccard�����
    ��
- �����rule-first�score-second��
    - ����������� `table_ratio>t1` → �����`code_ratio>t2`
    → ��/������
    - ����������
        - `heading_block` �� ~ `heading_ratio + list_ratio`�
        - `sentence_split_sliding` �� ~ `1 - heading_ratio`�
        ���� `p90`�
        - `table_batch` �� ~ `table_ratio`�
    - ���������� < ε��������������
- �����
    - �����`min(max(p50����, 150), 400)`�����������
    - ���`overlap = 0.15` �����������������������
    � `0.2`�
    - ��������>12 ���������������
- �����
    - ����������������/���
    - ������/������“�������/���”���
- ������
    - ������������� 3 ��+ �� `sentence_split_sliding`�

    - ������������200 ��15% overlap��

7.1 �������������

����$h = heading\_ratio,\ l = list\_ratio,\ t = table\_ratio,\ c = code\_ratio,\ p = p90\_para\_len$������$w_h, w_l, w_t, w_c, w_p$���

$$S_{heading\_block} = w_h \cdot h + w_l \cdot l - w_p \cdot \max\left(0, \frac{p - 300}{300}\right)$$

$$S_{sentence\_sliding} = 1 - h - 0.3 \cdot l - 0.2 \cdot t - 0.2 \cdot c + w_p \cdot \min\left(1, \frac{p}{400}\right)$$

$$S_{table\_batch} = w_t \cdot t$$

$$S_{code/log} = w_c \cdot c$$

- �������$w_h = 0.6,\ w_l = 0.4,\ w_t = 0.8,\ w_c = 0.8,\ w_p = 0.3$�
- �������������� $t > t_1$�$c > c_1$������ $\arg\max S$�������� $< \epsilon$��������������

7.2 ���

```
def recommend(file, k=6, emit_candidates=False):
    probes = sample_probes(file, k=k)              # �/�/� + ����
    features = profile(probes)                     # ��/��/����

    # ����
    if features.table_ratio > t1:
        strategy = "table_batch"
    elif features.code_ratio > t2:
        strategy = "code_log_block"
    else:
        # ��
        s_heading = w_h*features.heading_ratio + w_l*features.list_ratio - w_p*max(0, (fea
        s_sentence = 1 - features.heading_ratio - 0.3*features.list_ratio - 0.2*features.t
        s_table = w_t*features.table_ratio
        s_code = w_c*features.code_ratio
        scores = {
            "heading_block_length_split": s_heading,
            "sentence_split_sliding": s_sentence,
            "table_batch": s_table,
            "code_log_block": s_code,
        }
        strategy = argmax(scores)

    params = estimate_params(features, strategy)   # ����/overlap/���

    return {
        "strategy_id": strategy,
        "params": params,
        "candidates": scores if emit_candidates else None,
        "profile": features"
    }
```

8. ��������

- ������doc_id�source_format�page/slide�section_path�strategy_id�probe_sample_info�chunk_rule�

- **��**�
    - ��������������/�������������
    - ���������� vs ����������������
    - ��������/���
- **��/��**���������������rule hit�����������������
  �����

9. �����

- ���������/�/��������������������������
- ��������������/�������������page ranges��
- �����conversion ����������� source ��������������
  ��